On Training Bi-directional Neural Network Language Model with Noise Contrastive Estimation

Tianxing He, Yu Zhang, Jasha Droppo, and Kai Yu
cloudygoose@sjtu.edu.cn, yzhang87@csail.mit.edu, jdroppo@microsoft.com, kai.yu@sjtu.edu.cn

Overview
• Motivation: MLE is not suitable for training bi-directional neural network language model.
• Approach: Use sentence-level NCE to achieve sentence-level normalization.
• Experiments & Discussion: Our proposed model performs well on a sanity pseudo PPL check, but unfortunately, it did not outperform our uni-directional baselines.

Background: Recurrent Neural Network Language Model
• RNNLM encodes all history with recurrent connections.

Difficulty of Training Bi-directional Language Model
The definition of uni-directional lm ensures its sentence-level normalization, which enables us to apply MLE framework.

\[ P(W) = \prod_{i} P(w_i | w_{1..i-1}) \]
\[ \sum_{W} P_L(M|W) = 1 \]

However a bi-directional lm doesn’t satisfy that condition. For example:

\[ P(w_i | w_{1..i-1}, i+1..N) \]

Model Formulation
In this work, \( P(W) \) consists of the product of word-level scores (similar to uni-directional LM) and a learned normalization scalar \( c \), required by the NCE framework to ensure normalization.

\[ v_i = g(W_i | h_{i-1}, h_{i+1}, \beta) \quad \text{One-hot representation} \]
\[ h_i = tanh(W_i h_{i-1} + W_h h_i + b) \quad \text{Gated Recurrent Unit(GRU)} \]
\[ f_i(W) = \sum_{w_i \in V} \alpha_{w_i}(w_i) \quad \text{Learned normalization constant} \]
\[ f'(V) = \prod_{i} f_i(W) \]
\[ f^{NCE}(W) = f'(W) \exp(c) \]

Training bi-NNLM with NCE
• (Noise Contrastive Estimation) NCE fits an unnormalized model to the data distribution by learning a normalization constant.

\[ J_{NCE} = E_{P_{data}(W)}[\log P(D = 1|W; \theta)] + k E_{P_{noise}(W)}[\log P(D = 0|W; \theta)] \]
\[ P(D = 1|W; \theta) = P_W^{NCE}(W) / (P_W^{NCE}(W) + k P_{noise}(W)) \]
\[ P(D = 0|W; \theta) = k P_{noise}(W) / (P_W^{NCE}(W) + k P_{noise}(W)) \]

Training & Implementation Details
Stochastic Gradient Descent (SGD) with learning rate (lr) decaying is used.
The SRILM toolkit is used to build 4-GRAM models as baselines.

Experiments on ptb-rescore task
To make our training time tolerable, we designed a task similar to “sentence completion” on the ptb dataset. The models are expected to assign higher sentence-level scores to the original sentence than the distorted sentences:

We state two major observations:
• The proposed NCE training for bi-directional GRULM out-performs MLE training.
• The performance can only be improved when the amount of noise samples grow exponentially.

Conclusions
Our proposed NCE training for bi-directional NNLM out-performed the MLE trained model, however, it did not outperform the uni-directional baselines. The reason maybe that sentence-level sampling space is too sparse for our sampling to cover.